

Dimensionality Reduction for Signal Decoding

Roman Isachenko

MIPT-UGA

young researchers workshop

Supervisor: Vadim Strijov

July 1, 2021

Dimensionality Reduction for Signal Decoding

The problem of investigation is a model selection in the input and target spaces of high hidden correlations.

Challenge

The target variable is a vector whose components are dependent. Heterogeneous spaces of input and target variables have significantly excessive dimensionality.

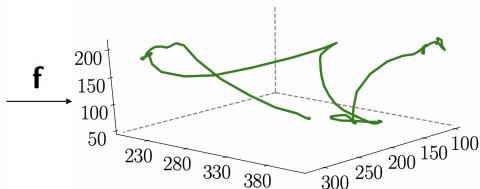
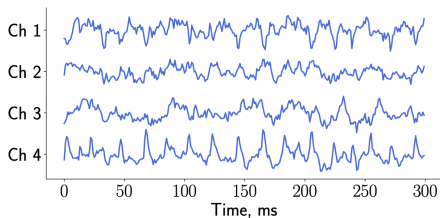
Goal

To build a model, which adequately describes the input and target spaces according to the observed multicorrelation in both spaces.

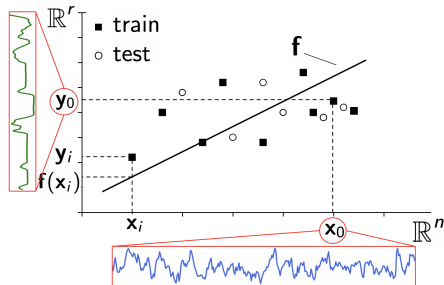
Solution

We propose the methods of dimensionality reduction, which project the input and target variables into the hidden low-dimensional space. We propose linear and nonlinear methods of building concordant models in high-dimensional spaces.

Concordance process in signal decoding task



Forecasting decoding model



Hidden space concordance

$$\begin{aligned}
 & \mathbf{x} \in \mathbb{R}^n \xrightarrow{\mathbf{f}} \mathbf{y} \in \mathbb{R}^r \\
 & \begin{array}{ccc}
 & \swarrow \mathbf{P} & \searrow \mathbf{Q} \\
 \mathbf{x} & & \mathbf{y} \\
 & \nwarrow \mathbf{W} & \nearrow \mathbf{C}
 \end{array} \\
 & \mathbf{t}, \mathbf{u} \in \mathbb{R}^\ell \\
 & \mathbf{x} = \mathbf{P}\mathbf{t} + \mathbf{e}_x \\
 & \mathbf{y} = \mathbf{Q}\mathbf{u} + \mathbf{e}_y \\
 & \text{cov}(\mathbf{t}, \mathbf{u}) \rightarrow \max_{\mathbf{P}, \mathbf{Q}}
 \end{aligned}$$

Signal decoding task

$\mathbf{Y} = \mathbf{F}(\mathbf{X}, \boldsymbol{\Theta}) + \mathbf{E}_y = \mathbf{X} \cdot \boldsymbol{\Theta}^\top + \mathbf{E}_y$ is a parametric model $\boldsymbol{\Theta} \in \mathbb{R}^{r \times n}$.

Decoding loss function

$$\mathcal{L}(f, \mathbf{X}, \mathbf{Y}) = \|\mathbf{Y} - \mathbf{F}(\mathbf{X}, \boldsymbol{\Theta})\|_2^2 = \left\| \underset{m \times r}{\mathbf{Y}} - \underset{m \times n}{\mathbf{X}} \cdot \underset{r \times n}{\boldsymbol{\Theta}^\top} \right\|_2^2 \rightarrow \min_{\boldsymbol{\Theta}}.$$

Projection to latent space

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\top + \mathbf{E}_x,$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\top + \mathbf{E}_y.$$

$$\begin{array}{ccc} \mathbf{X} \subset \mathbb{R}^n & \xrightarrow{\mathbf{f}} & \mathbf{Y} \subset \mathbb{R}^r \\ \mathbf{W} \updownarrow \mathbf{P} & & \mathbf{Q} \updownarrow \mathbf{C} \\ \mathbf{T} \subset \mathbb{R}^\ell & \xrightarrow{\mathbf{B}} & \mathbf{U} \subset \mathbb{R}^s \end{array}$$

Projection concordance

Link function couples images of input and target signals in latent space:

$$\mathbf{U} = \mathbf{T}\mathbf{B}, \quad \mathbf{B} = \text{diag}(\beta_k), \quad \beta_k = \mathbf{u}_k^\top \mathbf{t}_k / (\mathbf{t}_k^\top \mathbf{t}_k).$$

Decoding model:

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\top + \mathbf{E}_y \approx \mathbf{T}\mathbf{B}\mathbf{Q}^\top + \mathbf{E}_y = \mathbf{X}\mathbf{W}^*\mathbf{B}\mathbf{Q}^\top + \mathbf{E} = \mathbf{X}\boldsymbol{\Theta}^\top + \mathbf{E}_y,$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}^*, \quad \text{где } \mathbf{W}^* = \mathbf{W}(\mathbf{P}^\top \mathbf{W})^{-1}.$$

Concordance process in signal decoding task

The dimensionality of input and target spaces are highly redundant. We suggest to find manifolds of low dimensionality:

$$\begin{array}{ccc} \mathbb{X} \subset \mathbb{R}^n & \xrightarrow{\mathbf{f}} & \mathbb{Y} \subset \mathbb{R}^r \\ \varphi_e \updownarrow \varphi_d & & \psi_d \updownarrow \psi_e \\ \mathbb{T} \subset \mathbb{R}^\ell & \xrightarrow{\mathbf{h}} & \mathbb{U} \subset \mathbb{R}^s \end{array}$$

$\mathbb{T} \subset \mathbb{R}^\ell$ и $\mathbb{U} \subset \mathbb{R}^s$ are *latent spaces* for $\mathbb{X} \in \mathbb{R}^n$ ($\ell \leq n$) and $\mathbb{Y} \in \mathbb{R}^r$ ($s \leq r$), if there is *encoding function* $\varphi_e : \mathbb{X} \rightarrow \mathbb{T}$, $\psi_e : \mathbb{Y} \rightarrow \mathbb{U}$ and *decoding function* $\varphi_d : \mathbb{T} \rightarrow \mathbb{X}$, $\psi_d : \mathbb{U} \rightarrow \mathbb{Y}$:

for each $\mathbf{x} \in \mathbb{X}$ exists $\mathbf{t} \in \mathbb{T} : \varphi_d(\varphi_e(\mathbf{x})) = \varphi_d(\mathbf{t}) = \mathbf{x}$;

for each $\mathbf{y} \in \mathbb{Y}$ exists $\mathbf{u} \in \mathbb{U} : \psi_d(\psi_e(\mathbf{y})) = \psi_d(\mathbf{u}) = \mathbf{y}$.

The latent spaces \mathbb{T} и \mathbb{U} are called **concordant**, if there is a *link function* $\mathbf{h} : \mathbb{T} \rightarrow \mathbb{U}$:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) = \psi_d(\mathbf{h}(\varphi_e(\mathbf{x}))).$$

Projection concordance function

$$g : \mathbb{T} \times \mathbb{U} \rightarrow \mathbb{R}, \quad g(\mathbf{t}, \mathbf{u}) = g(\varphi_e(\mathbf{x}), \psi_e(\mathbf{y})) \rightarrow \max_{\varphi_e, \psi_e, \mathbf{h}}$$

Latent space concordant function

Statement (Isachenko, 2017)

Vectors \mathbf{t}_k и \mathbf{u}_k obtained by the iterative procedure:

$$\mathbf{t}_k := \frac{\mathbf{X}_k \mathbf{w}_k}{\|\mathbf{w}_k\|}, \quad \mathbf{w}_k := \mathbf{X}_k^\top \mathbf{u}_{k-1} / (\mathbf{u}_{k-1}^\top \mathbf{u}_{k-1});$$
$$\mathbf{u}_k := \frac{\mathbf{Y}_k \mathbf{c}_k}{\|\mathbf{c}_k\|}, \quad \mathbf{c}_k := \mathbf{Y}_k^\top \mathbf{t}_k / (\mathbf{t}_k^\top \mathbf{t}_k).$$

have largest covariance $\text{cov}(\mathbf{t}, \mathbf{u})$.

Theorem (Isachenko, 2017)

In the case of a linear encoding function $\varphi_e(\mathbf{T}) = \mathbf{T}\mathbf{P}^\top$, a linear decoding function $\psi_e(\mathbf{U}) = \mathbf{U}\mathbf{Q}^\top$ and a concordance function $g(\mathbf{t}, \mathbf{u}) = \text{cov}(\mathbf{t}, \mathbf{u})$ the model parameters

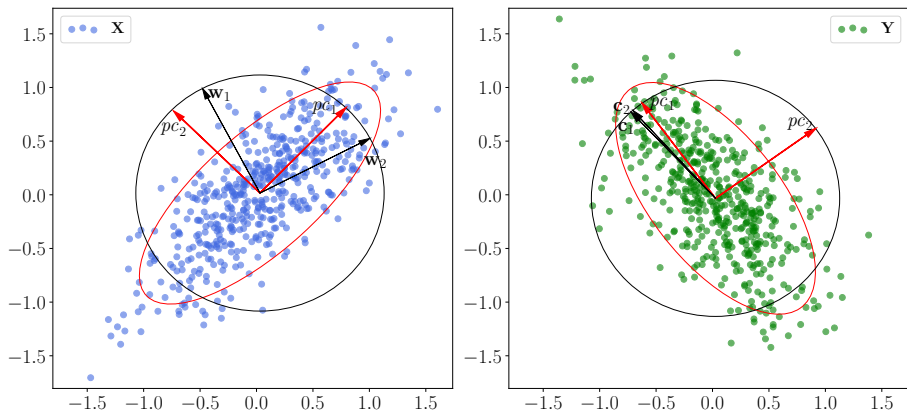
$$\Theta = \mathbf{W}(\mathbf{P}^\top \mathbf{W})^{-1} \mathbf{B} \mathbf{Q}^\top$$

are optimal for the model $\mathbf{F}(\mathbf{X}, \Theta)$.

Example of hidden space concordant projection

The input variables $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma})$.

The target variables \mathbf{y}_i linearly depend on pc_2 and does not depend on pc_1 .



Projection concordance of matrices \mathbf{X} and \mathbf{Y} find optimal hidden representation, deviating vectors \mathbf{w}_k and \mathbf{c}_k from principal components.

Superposition of signal decoding models

Let $\mathbf{f}_1(\mathbf{x}_1, \Theta_1)$, $\mathbf{f}_2(\mathbf{x}_2, \Theta_2)$ be linear signal decoding models

Statement (isachenko, 2021)

If the decoding model is an additive superposition of linear models:

$$\mathbf{y} = \mathbf{f}_1(\mathbf{x}_1, \Theta_1) + \mathbf{f}_2(\mathbf{x}_2, \Theta_2) + \varepsilon_{\mathbf{y}} = \Theta_1 \mathbf{x}_1 + \Theta_2 \mathbf{x}_2 + \varepsilon_{\mathbf{y}},$$

then the optimal parameters are

$$\Theta_1 = (\mathbf{X}_1^T \mathbf{M}_{\mathbf{X}_2} \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{M}_{\mathbf{X}_2} \mathbf{Y},$$

$$\Theta_2 = (\mathbf{X}_2^T \mathbf{M}_{\mathbf{X}_1} \mathbf{X}_2)^{-1} \mathbf{X}_2^T \mathbf{M}_{\mathbf{X}_1} \mathbf{Y},$$

where $\mathbf{M}_{\mathbf{X}_1} = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T$, $\mathbf{M}_{\mathbf{X}_2} = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2^T \mathbf{X}_2)^{-1} \mathbf{X}_2^T$.

Theorem (Isachenko, 2021)

If $\text{span}(\mathbf{X}_1) \neq \text{span}(\mathbf{X}_2)$, then the error of the additive superposition of linear decoding models is less or equal the error of each of the models:

$$\mathcal{L}_{\text{sup}}(\Theta_1^*, \Theta_2^*, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) \leq \mathcal{L}(\Theta_1, \mathbf{X}_1, \mathbf{Y}),$$

$$\mathcal{L}_{\text{sup}}(\Theta_1^*, \Theta_2^*, \mathbf{X}_1, \mathbf{X}_2, \mathbf{Y}) \leq \mathcal{L}(\Theta_2, \mathbf{X}_2, \mathbf{Y}).$$

Nonlinear methods of hidden space concordance

Let encoding and decoding functions be deep neural networks:

$$\mathbf{T} = \varphi_e(\mathbf{X}) = \mathbf{W}_x^L \sigma(\dots \sigma(\mathbf{W}_x^2 \sigma(\mathbf{X} \mathbf{W}_x^1)) \dots)$$

$$\mathbf{U} = \psi_e(\mathbf{Y}) = \mathbf{W}_y^L \sigma(\dots \sigma(\mathbf{W}_y^2 \sigma(\mathbf{Y} \mathbf{W}_y^1)) \dots)$$

$$\mathbf{X} = \varphi_d(\mathbf{T}) = \mathbf{W}_t^L \sigma(\dots \sigma(\mathbf{W}_t^2 \sigma(\mathbf{T} \mathbf{W}_t^1)) \dots)$$

$$\mathbf{Y} = \psi_d(\mathbf{U}) = \mathbf{W}_u^L \sigma(\dots \sigma(\mathbf{W}_u^2 \sigma(\mathbf{U} \mathbf{W}_u^1)) \dots)$$

Projection concordance

$$g(\mathbf{T}, \mathbf{U}) \rightarrow \max_{\mathbf{W}} \quad \mathbf{W} = \{\mathbf{W}_x^i, \mathbf{W}_y^i, \mathbf{W}_t^i, \mathbf{W}_u^i\}_{i=1}^L.$$

The gradient of the concordance function $g(\mathbf{t}, \mathbf{u}) = \text{corr}(\mathbf{t}, \mathbf{u})$ is

$$\frac{\partial g(\mathbf{t}, \mathbf{u})}{\partial \mathbf{t}} = \frac{1}{\ell - 1} \left(\boldsymbol{\Sigma}_1^{-1/2} \mathbf{U} \mathbf{V}^T \boldsymbol{\Sigma}_2^{-1/2} \mathbf{U} - \boldsymbol{\Sigma}_1^{-1/2} \mathbf{U} \mathbf{D} \mathbf{V}^T \boldsymbol{\Sigma}_1^{-1/2} \right),$$

where $\mathbf{U}, \mathbf{D}, \mathbf{V} = \text{SVD}(\boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_2^{-1/2}$, $\boldsymbol{\Sigma}_1 = \frac{1}{\ell-1} \mathbf{T} \mathbf{T}^T$,
 $\boldsymbol{\Sigma}_2 = \frac{1}{\ell-1} \mathbf{U} \mathbf{U}^T$, $\boldsymbol{\Sigma}_{12} = \frac{1}{\ell-1} \mathbf{T} \mathbf{U}^T$.

Feature selection for signal decoding task

$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ is the input variable matrix; $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_r] \in \mathbb{R}^{m \times r}$ is the target variable matrix.

The goal of feature selection is to find $\mathbf{a} = \{0, 1\}^n$, which components are the indicators of selected features.

Feature selection loss function

$$\mathbf{z} = \arg \min_{\mathbf{z}' \in [0,1]^n} S(\mathbf{z}', \mathbf{X}, \mathbf{Y}), \quad a_j = [z_j > \tau].$$

Loss function of decoding model

$$\mathcal{L}(\Theta_{\mathbf{a}}, \mathbf{X}_{\mathbf{a}}, \mathbf{Y}) = \left\| \mathbf{Y} - \mathbf{X}_{\mathbf{a}} \Theta_{\mathbf{a}}^T \right\|_2^2 \rightarrow \min_{\Theta_{\mathbf{a}}}, \quad \mathbf{X}_{\mathbf{a}} = \{\mathbf{x}_j : a_j = 1, j = 1, \dots, n\}$$

Quadratic programming feature selection

$$S(\mathbf{z}, \mathbf{X}, \mathbf{y}) = (1 - \alpha) \cdot \underbrace{\mathbf{z}^T \mathbf{Q} \mathbf{z}}_{\text{Sim}(\mathbf{X})} - \alpha \cdot \underbrace{\mathbf{b}^T \mathbf{z}}_{\text{Rel}(\mathbf{X}, \mathbf{y})} \rightarrow \min_{\substack{\mathbf{z} \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{z} = 1}}.$$

$\mathbf{z} \in [0, 1]^n$ is a feature importance score;

$\mathbf{Q} = [|\text{corr}(\mathbf{x}_i, \mathbf{x}_j)|]_{i,j=1}^n \in \mathbb{R}^{n \times n}$ is a pairwise feature similarity matrix;

$\mathbf{b} = [|\text{corr}(\mathbf{x}_i, \mathbf{y})|]_{i=1}^n \in \mathbb{R}^n$ is a vector of feature relevancies.

Quadratic programming feature selection

$$S(\mathbf{z}, \mathbf{X}, \boldsymbol{\nu}) = (1 - \alpha) \cdot \underbrace{\mathbf{z}^T \mathbf{Q} \mathbf{z}}_{\text{Sim}(\mathbf{X})} - \alpha \cdot \underbrace{\mathbf{b}^T \mathbf{z}}_{\text{Rel}(\mathbf{X}, \boldsymbol{\nu})} \rightarrow \min_{\substack{\mathbf{z} \geq 0_n \\ \mathbf{1}_n^T \mathbf{z} = 1}} .$$

Theorem (Isachenko, 2018)

Let the pairwise feature similarity matrix $\hat{\mathbf{Q}}$ is obtained by semidefinite relaxation of initial matrix \mathbf{Q} :

$$\hat{\mathbf{Q}} = \mathbf{Q} - \lambda_{\min}(\mathbf{Q})\mathbf{I}.$$

Then quadratic programming feature selection task has a unique global minimum.

Relevance aggregation (RelAgg)

$$\mathbf{b} = [|\text{corr}(\chi_i, \boldsymbol{\nu})|]_{i=1}^n \rightarrow \mathbf{b} = \left[\sum_{k=1}^r |\text{corr}(\chi_i, \boldsymbol{\nu}_k)| \right]_{i=1}^n .$$

Drawback: dependencies in the target space of the matrix \mathbf{Y} are not taking into account.

Feature selection in signal decoding task

Symmetric Importance (SymImp)

Let penalize target vectors by introducing the term $\text{Sim}(\mathbf{Y})$:

$$S(\mathbf{z}, \mathbf{X}, \mathbf{Y}) = \underbrace{\alpha_1 \cdot \mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \underbrace{\alpha_2 \cdot \mathbf{z}_x^T \mathbf{B} \mathbf{z}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} + \underbrace{\alpha_3 \cdot \mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} \rightarrow \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{z}_x = 1 \\ \mathbf{z}_y \geq \mathbf{0}_r, \mathbf{1}_r^T \mathbf{z}_y = 1}},$$

$$\mathbf{Q}_x = [|\text{corr}(\chi_i, \chi_j)|]_{i,j=1}^n, \quad \mathbf{Q}_y = [|\text{corr}(\nu_i, \nu_j)|]_{i,j=1}^r, \quad \mathbf{B} = [|\text{corr}(\chi_i, \nu_j)|]_{\substack{i=1,\dots,n, \\ j=1,\dots,r}}$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1, \quad \alpha_i \geq 0.$$

SymImp penalizes correlated target vectors, which are less explained by features.

Assymetric Importance (AsymImp)

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \underbrace{\alpha_2 \cdot (\mathbf{z}_x^T \mathbf{B} \mathbf{z}_y - \mathbf{b}^T \mathbf{z}_y)}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} + \underbrace{\alpha_3 \cdot \mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} \rightarrow \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{z}_x = 1 \\ \mathbf{z}_y \geq \mathbf{0}_r, \mathbf{1}_r^T \mathbf{z}_y = 1}}.$$

In the case of $b_j = \max_{i=1,\dots,n} [\mathbf{B}]_{i,j}$ the components of \mathbf{z}_y in $\text{Rel}(\mathbf{X}, \mathbf{Y})$ are non-negative.

Feature selection in signal decoding task

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^T \mathbf{B} \mathbf{z}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} \rightarrow \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{z}_x = 1}} ; \quad \alpha_3 \cdot \underbrace{\mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{z}_x^T \mathbf{B} \mathbf{z}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} \rightarrow \min_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{z}_y = 1}} .$$

MinMax or MaxMin

$$S(\mathbf{z}, \mathbf{X}, \mathbf{Y}) = \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{z}_x = 1}} \max_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{z}_y = 1}} \left(\text{or } \max_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{z}_y = 1}} \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{z}_x = 1}} \right) \left[\alpha_1 \cdot \underbrace{\mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^T \mathbf{B} \mathbf{z}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} - \alpha_3 \cdot \underbrace{\mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} \right].$$

Theorem (Isachenko, 2018)

In the case of positive definite matrices \mathbf{Q}_x and \mathbf{Q}_y , minmax and maxmin problems have the same value of the loss function $S(\mathbf{z}, \mathbf{X}, \mathbf{Y})$

Theorem (Isachenko, 2018)

Minmax task is equivalent to the quadratic programming task with $n + r + 1$ variables.

Semidefinite relaxation by shifting the matrix spectrum makes the task convex.

Generalization of the proposed feature selection methods

Theorem (Isachenko, 2018)

In the case of the scalar target variable $r = 1$, the proposed feature selection methods (SymImp, MinMax, MaxMin, AsymImp) coincide with the original problem of minimizing the loss function $S(\mathbf{z}, \mathbf{X}, \mathbf{Y})$.

Method	Criterion	Loss function $S(\mathbf{z}, \mathbf{X}, \mathbf{Y})$
RelAgg	$\min [\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X}, \mathbf{Y})]$	$\min_{\mathbf{z}_x} [(1 - \alpha) \cdot \mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x - \alpha \cdot \mathbf{z}_x^T \mathbf{B} \mathbf{1}_r]$
SymImp	$\min [\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X}, \mathbf{Y}) + \text{Sim}(\mathbf{Y})]$	$\min_{\mathbf{z}_x, \mathbf{z}_y} [\alpha_1 \cdot \mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^T \mathbf{B} \mathbf{z}_y + \alpha_3 \cdot \mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y]$
MinMax	$\min [\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X}, \mathbf{Y})]$ $\max [\text{Rel}(\mathbf{X}, \mathbf{Y}) + \text{Sim}(\mathbf{Y})]$	$\min_{\mathbf{z}_x} \max_{\mathbf{z}_y} [\alpha_1 \cdot \mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^T \mathbf{B} \mathbf{z}_y - \alpha_3 \cdot \mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y]$
AsymImp	$\min [\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X}, \mathbf{Y})]$ $\max [\text{Rel}(\mathbf{X}, \mathbf{Y}) + \text{Sim}(\mathbf{Y})]$	$\min_{\mathbf{z}_x, \mathbf{z}_y} [\alpha_1 \mathbf{z}_x^T \mathbf{Q}_x \mathbf{z}_x - \alpha_2 (\mathbf{z}_x^T \mathbf{B} \mathbf{z}_y - \mathbf{b}^T \mathbf{z}_y) + \alpha_3 \mathbf{z}_y^T \mathbf{Q}_y \mathbf{z}_y]$

Optimization of nonlinear models using quadratic programming

Let call the parameter θ_j of the model $f(\mathbf{x}, \theta)$ **active**, if $\mathbf{J}^T(\mathbf{f}(\mathbf{x}, \theta) - \mathbf{y}) \neq 0$.

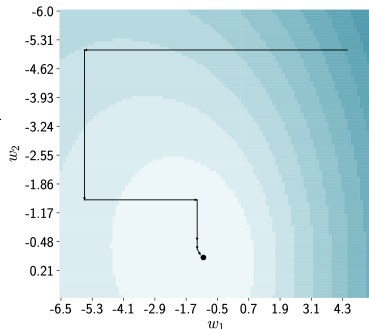
Theorem (Isachenko, 2018)

For differentiable model $\mathbf{f}(\mathbf{x}, \theta)$ the problem of selecting the model active parameters by the quadratic programming in the case of Newton method takes the form:

$$S(\mathbf{z}, \mathbf{X}, \nu) = (1-\alpha) \cdot \underbrace{\mathbf{z}^T \mathbf{Q} \mathbf{z}}_{\text{Sim}(\mathbf{J})} - \alpha \cdot \underbrace{\mathbf{b}^T \mathbf{z}}_{\text{Rel}(\mathbf{J}, \mathbf{f}-\mathbf{y})} \rightarrow \min_{\substack{\mathbf{z} \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{z} = 1}}$$

where

$$\mathbf{J} = \begin{pmatrix} \frac{\partial f(\mathbf{x}_1, \theta)}{\partial \theta_1} & \cdots & \frac{\partial f(\mathbf{x}_1, \theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{x}_m, \theta)}{\partial \theta_1} & \cdots & \frac{\partial f(\mathbf{x}_m, \theta)}{\partial \theta_p} \end{pmatrix}.$$



Quality criteria for solving the decoding task

Normalized RMSE

Forecasting quality:

$$\text{sRMSE}(\mathbf{Y}, \hat{\mathbf{Y}}_{\mathbf{a}}) = \sqrt{\frac{\text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}}_{\mathbf{a}})}{\text{MSE}(\mathbf{Y}, \bar{\mathbf{Y}})}} = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{\mathbf{a}}\|_2}{\|\mathbf{Y} - \bar{\mathbf{Y}}\|_2}, \quad \text{где} \quad \hat{\mathbf{Y}}_{\mathbf{a}} = \mathbf{X}_{\mathbf{a}} \boldsymbol{\Theta}_{\mathbf{a}}^{\top}.$$

$\bar{\mathbf{Y}}$ — constant forecast.

Multicorrelation

The mean value of coefficient of multiple correlation:

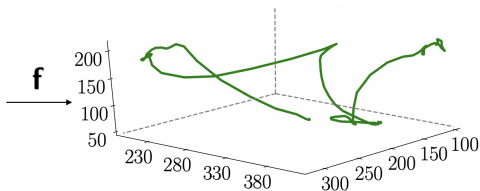
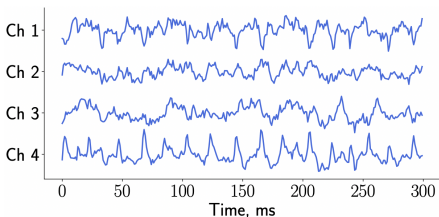
$$R^2 = \frac{1}{r} \text{tr} \left(\mathbf{C}^{\top} \mathbf{R}^{-1} \mathbf{C} \right), \quad \mathbf{C} = [\text{corr}(\chi_i, \nu_j)]_{\substack{i=1, \dots, n \\ j=1, \dots, r}}, \quad \mathbf{R} = [\text{corr}(\chi_i, \chi_j)]_{i,j=1}^n.$$

Bayesian information criteria

Trade-off of forecast quality and model complexity $\|\mathbf{a}\|_0$:

$$\text{BIC} = m \ln \left(\text{MSE}(\mathbf{Y}, \hat{\mathbf{Y}}_{\mathbf{a}}) \right) + \|\mathbf{a}\|_0 \cdot \log m.$$

Electrocorticogram signal decoding task



There are given:

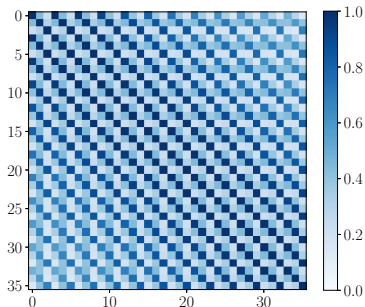
$\mathbf{X} \in \mathbb{R}^{m \times (32 \cdot 27)}$ – ECoG signals,

$\mathbf{Y} \in \mathbb{R}^{m \times 3k}$ – hand trajectory, where

$$\mathbf{Y} = \begin{pmatrix} x_1 & y_1 & z_1 & \dots & x_k & y_k & z_k \\ x_2 & y_2 & z_2 & \dots & x_{k+1} & y_{k+1} & z_{k+1} \\ \dots & \dots & \dots & & \dots & \dots & \dots \\ x_m & y_m & z_m & \dots & x_{m+k} & y_{m+k} & z_{m+k} \end{pmatrix}.$$

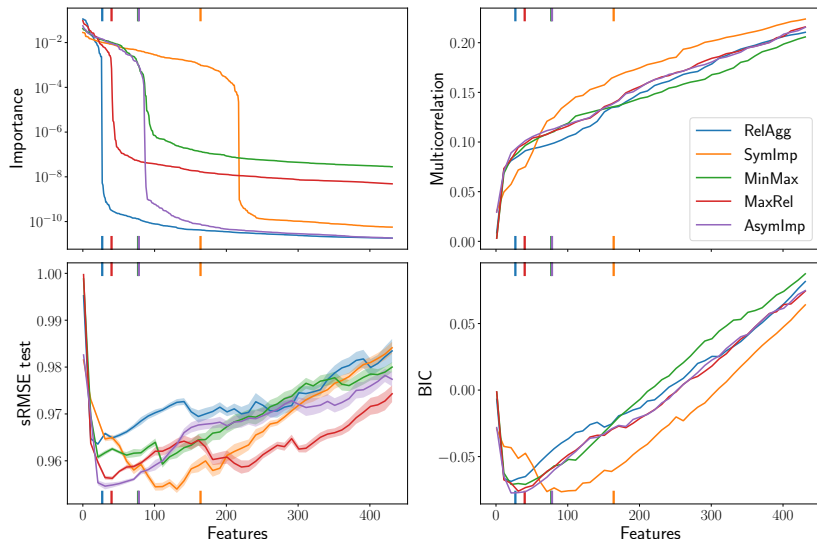
Columns of matrix \mathbf{Y} are highly correlated in the time axis.

<http://neurotycho.org>



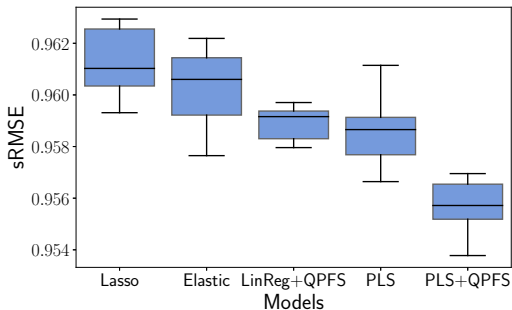
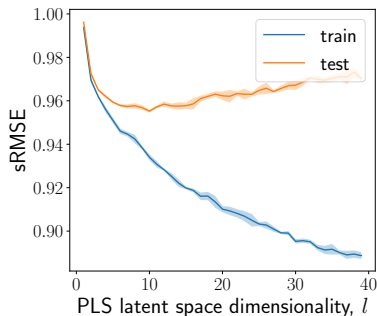
Correlation matrix \mathbf{Y}

Analysis of the proposed feature selection methods



The proposed methods for model selection have the smaller error in relation to the basic method.

Comparison of PLS method with the feature selection methods



- ▶ The proposed feature selection methods achieve a lower error compared to the basic Lasso and Elastic algorithms.
- ▶ PLS shows comparable quality with QPFS.
- ▶ A combination of two methods shows the best result.

Research results

1. We have investigated the problem of the signal dimensionality reduction in highly correlated spaces. We have proposed signal decoding methods that take into account the dependencies both in the input and in the target signal spaces.
2. We have proved the optimality theorems for the proposed methods of signal decoding. The proposed methods select consistent models in the case of excessive dimensionality of the data description.
3. We have proposed feature selection methods that take into account dependencies in both the input and target spaces. The proposed methods give stable and adequate solutions in high-dimensional spaces.
4. Nonlinear methods for hidden space concordance for complex structured target variable are proposed.
5. We have proposed the number of forecasting models for heterogeneous sets of signals that are suitable for the task of modelling brain computer interfaces.

Publication list

Papers

1. Isachenko R., Strijov V. Quadratic Programming Feature Selection for Multicorrelated Signal Decoding with Partial Least Squares *Expert Systems with Applications*, 2021, на рецензировании.
2. Исаченко Р.В., Яушев Ф.Р., Стрижов В.В. Модели согласования скрытого пространства в задаче прогнозирования // Системы и средства информатики, 31(1), 2021.
3. Isachenko R., Vladimirova M., Strijov V. Dimensionality Reduction for Time Series Decoding and Forecasting Problems. *DEStech Transactions on Computer Science and Engineering*, optim, 2018.
4. Isachenko R., Strijov V. Quadratic programming optimization for Newton method. *Lobachevskii Journal of Mathematics*, 39(9), 2018.
5. Isachenko R. et al. Feature Generation for Physical Activity Classification. *Artificial Intelligence and Decision Making*, 3, 2018.
6. Исаченко Р.В., Стрижов В. В. Метрическое обучение в задачах мультиклассовой классификации временных рядов *Информатика и её применения*, 10(2), 2016.

Conferences

1. Intelligent Data Processing Conference, 2020, Снижение размерности в задаче декодирования временных рядов.
2. Intelligent Data Processing Conference, 2018, Dimensionality reduction for multicorrelated signal decoding with projections to latent space.
3. Математические методы распознавания образов, 2017. Локальные модели для классификации объектов сложной структуры.
4. Intelligent Data Processing Conference, 2016. Multimodel forecasting multiscale time series in internet of things.
5. Ломоносов, 2016. Метрическое обучение в задачах мультиклассовой классификации временных рядов.